Sleep Medicine 15 (2014) 422–429

Contents lists available at ScienceDirect

Sleep Medicine

journal homepage: www.elsevier.com/locate/sleep

Original Article

Criterion validity of the Pittsburgh Sleep Quality Index and Epworth Sleepiness Scale for the diagnosis of sleep disorders

Takeshi Nishiyama [a,*], Tomoki Mizuno [b], Masayo Kojima [c], Sadao Suzuki [c], Tsuyoshi Kitajima [d], Kayoko Bhardwaj Ando [e,f], Shinichi Kuriyama [e,f], Meiho Nakayama [e,f]

[a] Department of Public Health, Aichi Medical University, Nagakute, Japan
[b] Nagoya City University Medical School, Nagoya, Japan
[c] Department of Public Health, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan
[d] Department of Psychiatry, Fujita Health University, Toyoake, Japan
[e] Department of Otolaryngology, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan
[f] Good Sleep Center, Nagoya City University Hospital, Nagoya, Japan

ARTICLE INFO

Article history:
Received 29 September 2013
Received in revised form 4 December 2013
Accepted 9 December 2013
Available online 22 February 2014

Keywords:
Pittsburgh Sleep Quality Index (PSQI)
Epworth Sleepiness Scale (ESS)
Obstructive sleep apnea (OSA)
Periodic limb movement disorder (PLMD)
Rapid eye movement sleep behavior disorder (RBD)
Narcolepsy

ABSTRACT

Objectives: (1) To examine criterion validity of the Pittsburgh Sleep Quality Index (PSQI) and Epworth Sleepiness Scale (ESS) using obstructive sleep apnea (OSA), periodic limb movement disorder (PLMD), rapid eye movement sleep behavior disorder (RBD), and narcolepsy as criterion standard. (2) To summarize the evidence for criterion validity of the ESS for the diagnosis of OSA by a meta-analysis that combines the current and previous studies. (3) To investigate the determinants of the PSQI and ESS scores.
Methods: The PSQI and ESS as well as the Hospital Anxiety and Depression Scale (HADS), which measures anxiety and depression levels, were administered to 367 patients consecutively referred to a sleep clinic. They underwent overnight polysomnography (PSG) and the multiple sleep latency test if narcolepsy was suspected.
Results: The area under the receiver operating characteristic curves for the ESS and PSQI (and its subscale) were <0.9, meaning that these questionnaires were not highly accurate for predicting the four sleep disorders. The meta-analysis found that the ESS had no value in identifying OSA. The variable that most strongly influenced PSQI or ESS scores was the HADS score.
Conclusion: The PSQI and ESS should no longer be used as a screening or diagnostic instrument for the four PSG-defined sleep disorders, especially in a low-risk population.
© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Many self-report questionnaires have been developed for measuring various aspects of sleep disturbance [1]. Two of the most widely used are the Pittsburgh Sleep Quality Index (PSQI) [2] and the Epworth Sleepiness Scale (ESS) [3]. The PSQI and the ESS were originally designed to measure sleep quality and subjective daytime sleepiness, respectively, but they were not designed to screen for a specific sleep disorder [2,3]. Nevertheless, they have been widely used in clinical settings, with the expectation that the PSQI can identify persons at high risk for insomnia or that the ESS can identify persons at high risk for OSA or narcolepsy by identifying persons with excessive daytime sleepiness [4,5].

Although a previous study examined the criterion validity of the PSQI using actigraphy, sleep diary, and a questionnaire of depression as criterion standard [6], there have been no studies using polysomnographic sleep abnormalities as criterion standard. On the other hand, several studies have been conducted to examine the criterion validity of the ESS for the diagnoses of OSA and narcolepsy [7–12]. For example, the original developer of the ESS used the case-control method in which a group of typically diagnosable patients with narcolepsy is compared with a group of unquestionably healthy subjects [5]. In this situation, the ability of the ESS to discriminate between the two groups may be overestimated. The failure to include an appropriately broad spectrum of diseased and non-diseased subjects in the study population may give falsely high sensitivity and specificity and falsely elevate the area under the receiver operating curve (AUC). This effect is known as spectrum bias [13].

* Corresponding author. Address: Department of Public Health, Aichi Medical University, 1-1 Yazako, Nagakute, Aichi 480-1195, Japan. Tel.: +81 561 62 3311; fax: +81 561 62 5270.
E-mail address: nishiyama@minos.ocn.ne.jp (T. Nishiyama).

http://dx.doi.org/10.1016/j.sleep.2013.12.015
1389-9457/© 2014 Elsevier B.V. All rights reserved.

Theoretically, a measure to distinguish individuals on an underlying dimension such as daytime sleepiness is totally distinct from a measure used as screening or diagnostic instrument to identify which specific individuals have a target condition such as OSA [14]. Therefore, in order to use the PSQI and ESS – which were originally developed for the former purpose as screening or diagnostic instrument for a specific sleep disorder – their criterion validity should be examined empirically. To our knowledge, however, there are no other studies of the criterion validities of the two measures that have used standard overnight polysomnography (PSG) for the diagnoses of sleep abnormalities, such as OSA, periodic limb movement disorder (PLMD), rapid eye movement (REM) sleep behavior disorder (RBD), and narcolepsy, with the exception of studies that have examined the use of the ESS for the diagnosis of OSA. Regarding the latter, two studies of the ESS were extracted in a previous meta-analysis of screening tests for OSA [15], but one of the two studies suffered from typical spectrum bias [3]. In a meta-analysis that was updated on the same topic [16,17], another study of the ESS was extracted [8]. Further, we identified additional articles covering this topic, as detailed below.

As revealed by the poor predictive ability of the ESS for diagnosing PSG-defined OSA in these studies [7–12], self-reported sleep measures, including the ESS and PSQI, are often inconsistent with objective sleep measures such as PSG indices. This discrepancy may, in part, be explained by the fact that self-reported sleep measures focus on subjective experiences of sleep that are not directly captured with PSG. Therefore, it seems a logical consequence that the ESS and PSQI are influenced by psychological symptoms, such as depression [6,18–21] and anxiety [21].

Therefore, the purpose of the current study is: (1) to create evidence for criterion validity of the PSQI and ESS for the diagnoses of OSA, PLMD, RBD, and narcolepsy; and (2) to summarize the evidence for criterion validity of the ESS for the diagnosis of OSA by a meta-analysis combining the current and previously published studies; and (3) to investigate the degree to which the sleep quality and sleepiness reported on the PSQI and ESS, respectively, reflect PSG indices and psychological distress.

2. Methods

2.1. Subjects

In all, 503 patients aged ⩾16 years, consecutively referred to the Good Sleep Center in Nagoya City University Hospital between April 2011 and December 2012, were recruited for this study. The patients presented with a variety of complaints, including night-time snoring, daytime sleepiness, insomnia, witnessed apnea during sleep, difficult or noisy breathing during sleep, leg cramps/pain, and abnormal oximetry. Of these 503 patients, 136 were excluded due to prior treatment for sleep disturbances, resulting in a sample size of 367 patients. After administration of Japanese versions of the PSQI [22], ESS [23], and the Hospital Anxiety and Depression Scale (HADS) [24], all of the subjects underwent at least one night of PSG assessment and, if narcolepsy was suspected, the multiple sleep latency test (MSLT) was administered on the day following PSG.

The present study was approved by the Nagoya City University Ethics Review Board Committee, and written informed consent was obtained from each participant prior to his or her participation in the study.

2.2. Measures

2.2.1. Pittsburgh Sleep Quality Index

The PSQI is a self-report measure of sleep quality [2]. It consists of 19 items plus a five-item rating made by a bed partner that is not included in scoring. Respondents indicate the amount of sleep they obtained and rate the extent to which various factors interfered with their sleep on a four-point Likert-type scale (0 = not at all, 3 = three or more times a week). These items yield scores on seven subscales: subjective sleep efficiency, sleep latency, sleep duration, sleep quality, sleep disturbance, sleep medication use, and daytime dysfunction due to sleepiness. The subscales yield a score from 0 to 3 and are summed to yield a total score ranging from 0 to 21, with higher total scores indicating poorer sleep quality.

2.2.2. Epworth Sleepiness Scale

The ESS is an eight-item self-report measure of excessive daytime sleepiness [3]. Respondents indicate, on a four-point Likert-type scale (0 = never, 3 = high chance), the likelihood that they will 'doze off or fall asleep' in eight different conditions (e.g. riding as a passenger in a car). The responses are summed to yield a total score from 0 to 24, with higher scores indicating greater sleepiness during common daily activities.

2.2.3. Hospital Anxiety and Depression Scale

The HADS is a 14-item self-report measure of psychological distress [25]. Respondents indicate the frequency of any symptoms on a four-point Likert-type scale. The HADS was designed to measure anxiety and depression (seven items for each subscale). The total score is the sum of the 14 items (ranging from 0 to 42), and for each subscale the score is the sum of the respective seven items (ranging from 0 to 21). The scale includes no items relating to sleep disturbances. Previous studies have reported moderate to strong correlations between the anxiety and depression subscales (0.49–0.74 in non-patient samples and 0.40–0.64 in physical and psychiatric patient samples) [26]. Therefore, we used only the HADS total score as an independent variable in the regression analyses described below.

2.3. Polysomnography

All study participants underwent full-night PSG, and a sleep medicine physician interpreted the results. A 12-channel montage was utilized to record electroencephalogram, electro-oculogram, electrocardiogram, chin and lower extremity electromyogram, naso-oral airflow, thoracic and abdominal effort, and oxygen saturation by pulse oximeter. All subjects were evaluated in an accredited sleep laboratory in sound-attenuated rooms, monitored by an infra-red camera. The records were scored according to the method of Rechtschaffen and Kales [27]. The EEG was divided into non-rapid eye movement (NREM) and rapid eye movement (REM) states. NREM is conventionally subdivided into four stages, 1–4. Sleep stages 3 and 4 were scored together as delta sleep, and movement time was scored as an arousal (2–15 s) or awakening (>15 s). The analyzed indices of sleep architecture were total sleep time (TST), sleep efficiency (TST divided by time in bed), percentage of sleep stages, and apnea–hypopnea index (AHI), defined as the average number of apneas and hypopneas per hour of sleep, where apnea was defined as a complete cessation of airflow lasting >10 s, and hypopnea was defined as ⩾30% airflow reduction lasting >10 s and ⩾4% desaturation [28]. Arousal index (AI) was scored as described by the American Sleep Disorders Association [29]. Periodic limb movement index (PLMI) was scored according to standard criteria [30]. Digital videotaping with sound recording was performed throughout the night.

2.4. Clinical expert diagnosis of sleep disorders

The diagnoses of RBD, PLMD, and narcolepsy were made by experienced sleep clinicians, who were blinded to the responses

on the two questionnaires, based on the case histories and PSGs according to the ICSD-2 criteria [31]. The diagnosis of OSA was assigned based on the criteria of AHI $\geqslant$ 5, and almost all apnea/hypopnea events were confirmed as obstructive by the presence of persistent respiratory effort despite an absence of airflow. The other sleep disturbances, except for the four sleep disorders, were described as 'other sleep disorders' in the study. This group included 38 patients with insomnia, a patient with hypersomnia, and a patient with restless leg syndrome. All subjects in this study had at least one sleep disorder.

2.5. Meta-analysis

A meta-analysis of studies regarding the predictive ability of the ESS for OSA was undertaken according to the PRISMA guidelines [32]. Two independent authors (T.N. and T.M.) performed a Medline search for English language articles from inception to January 2013, using the key words and MeSH terms: 'Epworth Sleepiness Scale or ESS' and 'obstructive sleep apnea or sleep apnea or OSA or sleep disordered breathing.' They also searched the references of the potentially eligible studies identified above. The search results were evaluated independently by the two authors to find the eligible articles for inclusion, based on the following criteria: (1) the diagnosis of OSA was made using standard overnight PSG as the 'gold standard', (2) the study population was comprised of adults aged $\geqslant$16 years, and (3) the study result was presented to allow the construction of a 2 $\times$ 2 contingency table. The methodological quality of each article was assessed independently by the two authors, based on the guidelines for meta-analysis evaluating diagnostic tests [33]. All disagreements were resolved by consensus. The full PRISMA checklist and diagram are provided as Supplementary Table 1 and Supplementary Fig. 1.

2.6. Data analysis

Differences in the descriptive characteristics between the five diagnostic groups were assessed using means and their corresponding 95% confidence intervals (95% CIs) using the bootstrap BCa method with 10,000 bootstrap samples [34]. To assess the criterion validities of the ESS and the PSQI and its subscales, receiver operating characteristics (ROC) analyses were conducted with the clinical diagnosis of each sleep disorder as a criterion standard. The area under the ROC curves (AUCs) and the 95% CIs were calculated by the non-parametric method [35]. When a high AUC for a diagnosis was noted, the sensitivities, specificities, positive likelihood ratios, and negative likelihood ratios for that diagnosis were estimated across all possible cut-off values within the range of the questionnaire.

To summarize study findings (including those from the current study) about the validity of the ESS for the diagnosis of OSA we plotted the diagnostic odds ratio (DOR) for each study, and calculated the pooled DOR using a random-effects model [36]. The DOR describes the odds of a positive test result in patients with OSA compared with the odds of a positive test result in those without OSA. For the diagnosis of OSA, we used the criteria for defining OSA chosen by the author of each paper, as these differed across studies (e.g. AHI $\geqslant$ 15 [9,10] and AHI $\geqslant$ 5 [8]). The amount of heterogeneity between studies was assessed using a Q-test based on the DerSimonian–Laird estimator. A summary ROC curve and the 95% CI were calculated using the Lehmann model [37]. Publication bias was assessed using a funnel plot of the standard error versus the effect size on a logarithmic scale, and a rank correlation test for funnel plot asymmetry was also performed [38].

Finally, univariate regression analysis was performed to examine the associations of the ESS and PSQI with objective and subject characteristics that could influence ESS or PSQI scores such as age, gender, body mass index (BMI, kg/m^2), the HADS score, TST, PLMI, AHI, and sleep efficiency. Variables identified in univariate analyses as significant at $P < 0.10$ were included in subsequent multivariate analyses to evaluate their relative influence on ESS or PSQI scores. A partial correlation statistic was used to describe the amount of total variation in the ESS or PSQI scores that could be explained by each predictor.

To compute AUCs with 95% CIs and bootstrap 95% CIs of a mean, the R packages pROC [39] and boot [40] were used, respectively. To perform the diagnostic meta-analysis, the R packages metaphor [41] and mada [42] were used. All analyses, except as otherwise noted, were performed using R version 2.15.0 for Windows [43].

3. Results

The diagnoses of sleep disorders in the symptomatic cohort, excluding the other sleep disorders group, which are mostly comprised of insomnia, substantially overlapped as shown in Fig. 1. Therefore, it is necessary to be careful when comparing each descriptive characteristic between sleep disorder groups. All we can test by examining the overlaps of 95% CIs is a comparison between the other sleep disorders group and the group with OSA, PLMD, RBD, or narcolepsy.

Descriptive characteristics for the study population are presented in Table 1. Many indices of PSG were different between the other sleep disorders group and the group with OSA, PLMD, or RBD, because the majority of patients with PLMD and RBD have comorbid OSA (Fig. 1). This general pattern was observed in the variables: age, BMI, stage I time (%), AI, snoring time (%), and PLMI; all of which were higher in the OSA, PLMD, or RBD group than in the other sleep disorders group. Further, the stage II (%) and stage III–IV times (%) were nearly aligned with this pattern of contrast, although exceptions occurred in the PLMD and RBD groups. On the other hand, the narcolepsy group had significantly greater stage REM time (%) and sleep efficiency (%) than the other sleep disorders groups. Scores for component 7 (daytime dysfunction subscale) of the PSQI were much higher in the narcolepsy group than in any other group, and were also significantly increased in the narcolepsy group compared to the other sleep disorders groups.

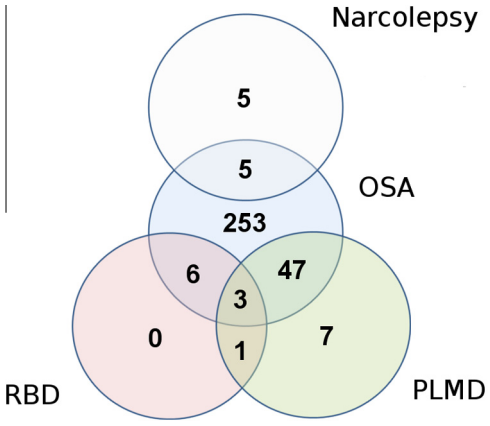

Fig. 1. Venn diagrams showing the diagnostic overlaps between obstructive sleep apnea (OSA), periodic limb movement disorder (PLMD), and rapid eye movement sleep behavior disorder (RBD). The numbers of subjects diagnosed with OSA, PLMD, RBD, and narcolepsy were 314, 58, 10, and 10, respectively. The numbers of subjects with more than one sleep disorder are represented in the overlapping portions of the circles, whereas those with one sleep disorder are represented in the non-overlapping portions of the circles. Note that narcolepsy did not overlap with these disorders.

Table 1
Sample demographics and the assessment measures used.[a]

	OSA	PLMD	RBD	Narcolepsy	Others[b]
No. of subjects[c]	314	58	10	10	40
Sex					
Female	96 (31%)	19 (33%)	2 (20%)	5 (50%)	29 (73%)
Male	218 (69%)	39 (67%)	8 (80%)	5 (50%)	11 (27%)
Age (years)	58.7 (57.0–60.3)	66.6 (63.4–69.4)	71.7 (68.2–76.3)	30.7 (23.6–43.9)	40.5 (35.5–45.9)
BMI (kg/m^2)	25.6 (25.1–26.0)	23.9 (23.1–24.7)	24.6 (23.1–26.3)	21.1 (19.6–23.2)	21.5 (20.5–22.6)
HADS	12.0 (11.2–12.7)	10.8 (9.3–12.7)	11.2 (9.2–13.7)	17.0 (12.8–21.3)	14.1 (11.9–17.0)
Sleep period time (min)	515.5 (507.1–523.5)	512.9 (494.5–528.2)	509.1 (473.4–550.0)	549.8 (504.0–580.4)	532.3 (499.9–554.1)
Total sleep time (min)	422.2 (411.8–432.2)	392.7 (368.5–414.3)	352.4 (276.2–412.4)	513.0 (476.5–544.1)	464.9 (432–492.2)
Stage I (%)	21.1 (19.7–22.8)	18.7 (16.3–21.8)	28.7 (24.0–37.8)	8.5 (5.3–12.7)	11.4 (9.1–15.9)
Stage II (%)	55.3 (53.9–56.6)	58.5 (55.2–61.5)	48.4 (37.7–56.6)	54.4 (50.7–58.8)	60.4 (57.4–63.5)
Stage III–IV (%)	4.4 (3.8–5.0)	4.5 (3.3–5.8)	5.9 (2.3–15.8)	12.6 (7.0–18.4)	8.5 (6.6–10.8)
Stage REM (%)	19.3 (18.4–20.1)	18.3 (16.6–20.4)	17.0 (10.7–22.0)	24.6 (22.2–27.5)	19.8 (17.6–21.6)
Sleep efficiency (%)	79.4 (77.6–80.9)	74.1 (69.5–78.0)	66.5 (52.5–76.3)	93.5 (91.0–95.6)	85.2 (81.3–88.2)
Arousal index	29.1 (27.2–31.3)	24.9 (22.6–27.4)	26.1 (21.0–32.9)	13.9 (9.9–18.4)	12.5 (11.0–14.6)
Snoring (%)[d]	34.3 (31.5–37.1)	27.6 (21.1–34.7)	38.1 (24.3–53.3)	13.6 (4.7–26.0)	12.2 (7.5–19.1)
AHI	30.7 (28.3–33.3)	21.4 (17.9–25.6)	26.5 (18.2–40.0)	6.0 (3.4–9.4)	2.5 (2.1–2.9)
PLMI	7.0 (5.6–9.0)	34.7 (30.6–40.9)	20.2 (9.1–42.9)	0.8 (0.3–1.8)	1.9 (1.0–3.3)
ESS	8.2 (7.6–8.7)	7.0 (5.8–8.2)	5.1 (3.8–6.7)	14.0 (10.6–16.4)	9.2 (7.6–11.1)
PSQI[e]	7.2 (6.8–7.6)	7.6 (6.6–8.6)	5.6 (4.4–7.8)	8.8 (7.1–11.1)	7.1 (6.0–8.2)
C1. Subjective sleep quality	1.6 (1.5–1.7)	1.7 (1.6–1.9)	1.4 (0.8–1.7)	1.9 (1.4–2.2)	1.5 (1.3–1.7)
C2. Sleep latency	1.1 (1.0–1.2)	1.3 (1.0–1.5)	0.9 (0.3–1.3)	1.3 (0.6–1.9)	1.3 (0.9–1.7)
C3. Sleep duration	1.4 (1.3–1.4)	1.3 (1.0–1.6)	1.5 (0.9–1.9)	1.4 (0.8–1.8)	1.1 (0.8–1.3)
C4. Habitual sleep efficiency	0.4 (0.3–0.5)	0.5 (0.3–0.8)	0.5 (0.0–1.3)	0.4 (0.0–1.2)	0.3 (0.1–0.6)
C5. Sleep disturbances	1.0 (1.0–1.1)	1.1 (0.9–1.2)	0.8 (0.3–0.9)	1.2 (0.7–1.5)	1.1 (0.8–1.2)
C6. Use of sleep medication	0.7 (0.5–0.8)	0.8 (0.5–1.2)	0.0 (–[f])	0.3 (0.0–0.9)	0.7 (0.3–1.1)
C7. Daytime dysfunction	1.1 (1.0–1.2)	0.8 (0.6–1.1)	0.5 (0.1–0.7)	2.3 (1.4–2.7)	1.1 (0.8–1.4)

OSA, obstructive sleep apnea; RBD, rapid eye movement sleep behavior disorder; PLMD, periodic limb movement disorder; BMI, body mass index; HADS, Hospital Anxiety and Depression Scale; REM, rapid eye movement; AHI, apnea–hypopnea index; PLMI, periodic limb movement index; ESS, Epworth Sleepiness Scale; PSQI, Pittsburgh Sleep Quality Index.
[a] Summarized as a mean (corresponding bootstrap 95% confidence interval) for continuous data and count (percentage) for categorical data.
[b] 'Other' refers to the sleep disorders other than OSA, PLMD, RBD, and narcolepsy.
[c] Because each sleep disorder group overlaps with other sleep disorder groups, the total number does not equal sample size (n = 367).
[d] C1–7 refers to each component score of the PSQI.
[e] Percentage of total sleep time spent snoring.
[f] Incomputable because all data = 0.

The ROC analyses for the ESS and the PSQI and its subscales for the four sleep disorders showed that neither (sub)scale was highly accurate for these sleep disorders, defined as AUC < 0.9 (Table 2; Fig. 2) [44]. The two most accurate are the component 7 of the PSQI and the ESS for the diagnosis of narcolepsy (AUC: 0.82 and 0.80, respectively) and even the two measures have limited utility for predicting the narcolepsy diagnosis. The sensitivities, specificities, positive likelihood ratios, and negative likelihood ratios across all possible cut-off values of the PSQI component 7 for the diagnosis of narcolepsy are presented in Table 3: that subscale shows the highest AUC among the scales and subscales examined. Using a cut-off of >1 resulted in relatively high sensitivity (90%) and low specificity (69%). The sensitivities, specificities, positive likelihood ratios, and negative likelihood ratios using cut-offs of >10, >4, and >1 for the ESS score, PSQI global, and PSQI component scores, respectively, for each sleep disorder are also presented in Supplementary Table 2, though these were not clinically meaningful.

Our literature search regarding predictive ability of the ESS for OSA revealed 856 Medline entries, of which 16 met the eligibility criteria for use in the meta-analysis. After a detailed review of these articles, six articles were selected for the final analysis [7–12]. Characteristics of the included studies are summarized in Table 4. Including the current study, seven studies were pooled to yield a total of 1280 subjects. The average prevalence of OSA was 56.1% (range, 45.7–78.1%). The pooled DOR was 0.82 (95% CI, 0.57–1.19; Fig. 3). Fig. 4 shows the summary ROC curve for these

Table 2
Area under the receiver operating characteristics curves (95% confidence intervals) for the ESS score and the PSQI total and component scores for the four sleep disorders examined.

	OSA	RBD	PLMD	Narcolepsy
ESS	0.47 (0.38–0.55)	0.68 (0.57–0.79)	0.58 (0.50–0.66)	0.80 (0.67–0.93)
PSQI	0.50 (0.41–0.58)	0.65 (0.49–0.80)	0.52 (0.44–0.60)	0.65 (0.51–0.79)
C1. Subjective sleep quality	0.53 (0.45–0.61)	0.57 (0.40–0.73)	0.55 (0.48–0.63)	0.60 (0.43–0.78)
C2. Sleep latency	0.45 (0.36–0.53)	0.46 (0.29–0.62)	0.56 (0.49–0.64)	0.55 (0.36–0.74)
C3. Sleep duration	0.56 (0.49–0.64)	0.56 (0.39–0.73)	0.49 (0.41–0.58)	0.52 (0.35–0.68)
C4. Habitual sleep efficiency	0.53 (0.47–0.58)	0.50 (0.36–0.65)	0.52 (0.45–0.58)	0.50 (0.36–0.63)
C5. Sleep disturbances	0.48 (0.40–0.56)	0.40 (0.29–0.51)	0.52 (0.46–0.58)	0.58 (0.41–0.74)
C6. Use of sleep medication	0.49 (0.43–0.56)	0.37 (0.35–0.40)	0.54 (0.47–0.60)	0.43 (0.32–0.53)
C7. Daytime dysfunction	0.48 (0.40–0.57)	0.68 (0.57–0.79)	0.41 (0.33–0.48)	0.82 (0.66–0.97)

ESS, Epworth Sleepiness Scale; PSQI, Pittsburgh Sleep Quality Index; OSA, obstructive sleep apnea; RBD, rapid eye movement sleep behavior disorder; PLMD, periodic limb movement disorder.
C1–7 refers to each component score of the PSQI.

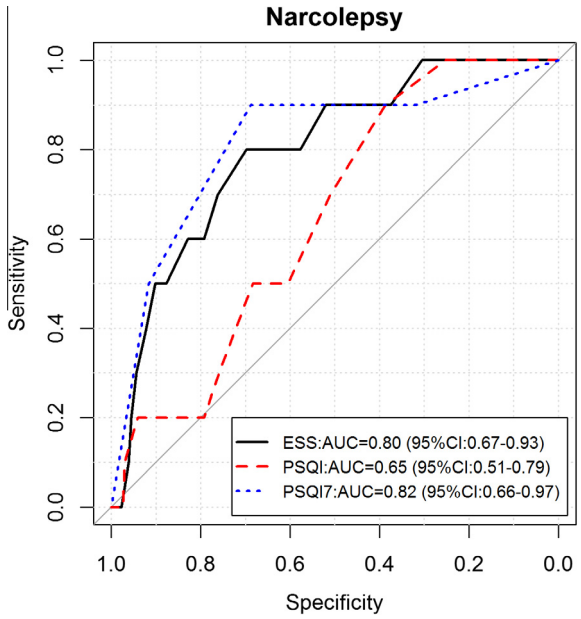

Fig. 2. Receiver operating characteristic (ROC) curves for the Epworth Sleepiness Scale (ESS) score, the Pittsburgh Sleep Quality Index (PSQI) total and component 7 score for the diagnosis of narcolepsy. Component 7 of the PSQI 7 and the ESS had areas under the ROC curves (AUCs) that were marginally acceptable for use in the diagnoses of narcolepsy; the AUC for the component 7 of the PSQI was 0.82 (95% CI, 0.66–0.97), and the AUC for the ESS was 0.80 (95% CI, 0.67–0.93). In contrast, the global PSQI score did not produce a clinically meaningful AUC for the diagnosis of narcolepsy (0.65; 95% CI, 0.51–0.79).

Table 3
Sensitivities, specificities, positive likelihood ratios, and negative likelihood ratios across all cut-off points for the PSQI component 7 score for the diagnosis of narcolepsy.

Cut-off	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR− (95% CI)
Score >0	0.90 (0.55–1.00)	0.32 (0.27–0.37)	1.33 (1.07–1.65)	0.31 (0.05–2.01)
Score >1	0.90 (0.55–1.00)	0.69 (0.64–0.73)	2.88 (2.22–3.72)	0.15 (0.02–0.94)
Score >2	0.50 (0.19–0.81)	0.91 (0.88–0.94)	5.77 (2.85–11.69)	0.55 (0.29–1.02)
Score >3	0.00 (0.00–0.31)	1.00 (0.98–1.00)	10.9 (0.47–252.3)	0.96 (0.84–1.09)

LR+, positive likelihood ratio; LR−, negative likelihood ratio; CI, confidence interval.

data. There was no statistically significant heterogeneity between the studies (Cochran's *Q*-statistics, $\chi^2 = 11.07$, $df = 7$, $P = 0.086$), and all of these data were used for the summary ROC analysis. The area under the summary ROC curve was 0.48 (95% CI, 0.41–0.56). A funnel plot of the inverse of the standard error versus

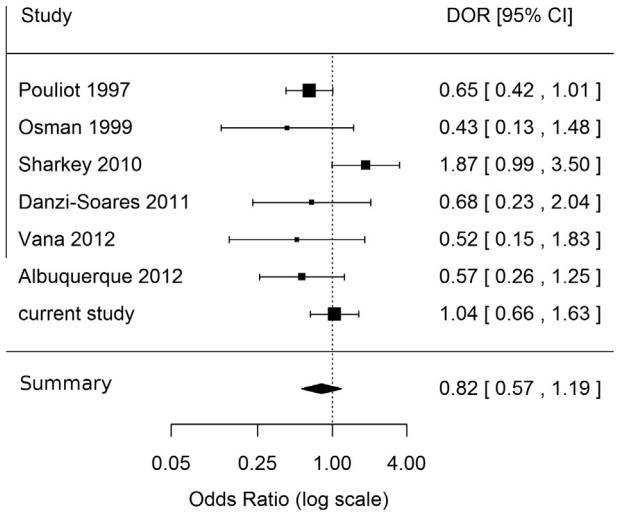

Fig. 3. Forest plot showing the results of seven studies examining the predictive ability of the Epworth Sleepiness Scale (ESS) for the diagnosis of obstructive sleep apnea (OSA). Forest plot with individual and pooled diagnostic odds ratios (DORs) were constructed with the sensitivity and specificity data of the ESS score for the diagnosis of OSA, using the definition of OSA diagnosis chosen by the investigators in each study. The pooled DORs were estimated using a random-effects model.

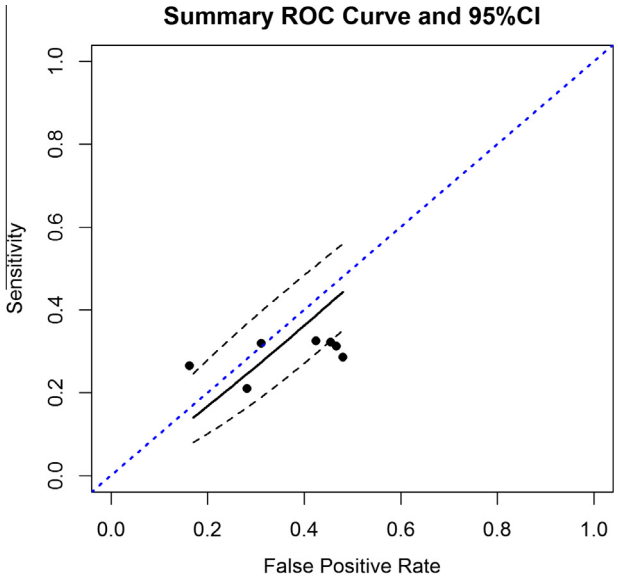

Fig. 4. Summary receiver operating characteristics (ROC) curve and its 95% confidence interval. Each black point represents the result of an individual study. The solid line is an estimate of the summary ROC curve, and the dotted lines are the 95% confidence interval of the curve. This curve is very close to the diagonal dashed line, which indicates that the ESS had poor diagnostic accuracy.

Table 4
Summary of studies included studies in the meta-analysis.

Study	Country	Settings	Sample size	Criteria for OSA[a]	Prevalence (95% CI)	Cut-off
Pouliot et al. [7]	Canada	Sleep clinic	354	AI ⩾ 20	0.46 (0.41–0.51)	ESS > 12
Osman et al. [8]	UK	Otolaryngological department	46	AHI ⩾ 5	0.46 (0.31–0.61)	ESS > 10
Sharkey et al. [9]	USA	Preoperative screening for bariatric surgery	245	AHI ⩾ 15 (5)	0.52 (0.46–0.59)	ESS > 11
Danzi-Soares et al. [10]	Brazil	Cardiovascular department	70	AHI ⩾ 15 (5)	0.54 (0.42–0.66)	ESS > 10
Vana et al. [11]	USA	Sleep clinic	47	AHI ⩾ 5	0.68 (0.53–0.81)	ESS > 10
Albuquerque et al. [12]	USA	Sleep clinic	151	AHI ⩾ 5 (10, 15, 30)	0.78 (0.71–0.84)	ESS > 10
Current study	Japan	Sleep clinic	367	AHI ⩾ 5	0.86 (0.82–0.89)	ESS > 10

OSA, obstructive sleep apnea; CI, confidence interval; ESS, Epworth Sleepiness Scale.
[a] Diagnostic criteria used for the meta-analysis are described, and other criteria are in parentheses when data for different criteria were available.

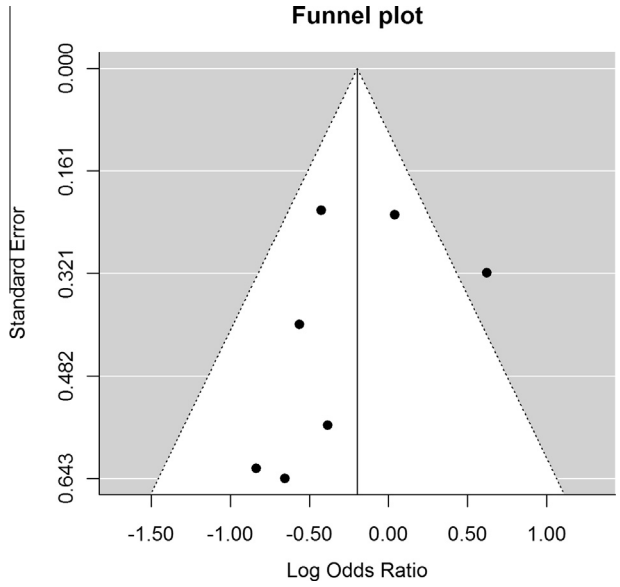

Fig. 5. Funnel plot showing a measure of variability (standard error) against the diagnostic odds ratios on a logarithmic scale. The points represent studies included in the meta-analysis. The studies are distributed evenly on either side of the vertical line, which indicates no clear evidence of publication bias.

the effect size showed no clear evidence of publication bias (Fig. 5). The rank correlation test produced a low Kendall's tau correlation coefficient of −0.05 ($P = 1.00$), demonstrating no clear evidence of publication bias.

Finally, we investigated the degrees to which the sleep quality and sleepiness reported on the PSQI and ESS, respectively, were influenced by PSG indices and psychological distress. Of the subject characteristics that may influence self-reported sleep quality and sleepiness and thus have been used in the previous studies [18,19,45], age, gender, BMI, the HADS score, TST, PLMI, AHI, and sleep efficiency were tested in univariate regression analyses with either PSQI or ESS scores as dependent variables. The variables screened by the univariate analyses for the PSQI were HADS score ($P < 0.001$), gender ($P < 0.001$), BMI ($P = 0.039$), sleep efficiency ($P = 0.009$), and AHI ($P = 0.0.005$); for the ESS, these variables were the HADS score ($P < 0.001$), age ($P < 0.001$), TST ($P < 0.001$), and sleep efficiency ($P < 0.001$). In subsequent multivariate analyses evaluating the relative influences of these screened variables on the ESS and PSQI, the HADS score remained a significant predictor of both questionnaires, whereas the significant associations of some predictors disappeared (Table 5). In these analyses, the partial r^2 of the HADS was much larger than that of other predictors, accounting for nearly all of the variation in the model's prediction of each score.

4. Discussion

Although the PSQI and ESS were not originally designed to screen for specific sleep disorders [2,3] and their criterion validity for sleep disorders has not been established in previous studies, they have been widely used in clinical settings to screen for sleep disorders, particularly the ESS for the diagnosis of OSA. Therefore, using PSG-defined sleep disorders as gold standard, we examined the criterion validity of the two self-reported questionnaires. As expected from the original intentions of the measures, our findings indicate significant limitations in identifying OSA, PMLD, and RBD using the two questionnaires. The ESS was demonstrated to have no value in identifying OSA with a high degree of certainty by meta-analysis of the current study and six previous studies that were extracted. At the maximum, the predictive abilities of both the PSQI daytime dysfunction subscale and the ESS for the diagnosis of narcolepsy proved to be moderately accurate in sleep clinic settings [44].

The current study was conducted in the sleep laboratory population. It is important to recognize that basic differences exist between the population in the sleep clinic and other medical settings. Patients are referred to a sleep clinic because of a perceived high risk for sleep disorders, whereas physicians use a screening tool in a lower-risk population. A more accurate screening test is of greater importance in the low-risk population than in the high-risk population that is seen in sleep clinic settings. For example, using the result of the PSQI daytime dysfunction subscale obtained in the current study (Table 3), a patient with a pre-test probability of 30% for narcolepsy who shows a PSQI daytime dysfunction score >1 has a post-test probability of 55% for the disorder, whereas a patient with a pre-test probability of 3% who shows a daytime dysfunction score >1 has a post-test probability of 5%. Clearly, a questionnaire with higher sensitivity might be essential for use in both low- and high-risk populations than the PSQI and ESS.

As such, many questionnaires have been developed to screen for specific sleep disorders. For example, the STOP questionnaire (snoring, tiredness during daytime, observed apnea, and high blood pressure), Berlin questionnaire, and American Society of Anesthesiologists checklist revealed AUCs of 0.77, 0.67, and 0.62, respectively, for the diagnosis of OSA, defined as AHI > 15 in consecutive patients from preoperative clinics [46]. The Mayo Sleep Questionnaire (MSQ) has an excellent predictive ability for RBD, with a sensitivity of 98% and a specificity of 74% in consecutive patients who mostly had dementia [47]. Though there is a spectrum bias inherent in the case–control method, the Ullanlinna Narcolepsy Scale (UNS) revealed a sensitivity of 100% and a specificity of 98.8% in various patient groups with narcolepsy, OSA, or depression [48].

Table 5
Results of multiple regression analysis for determinants of the PSQI and ESS scores.

	PSQI			ESS		
	Beta (95% CI)	Partial R^2	*P*-value	Beta (95% CI)	Partial R^2	*P*-value
HADS	0.16 (0.10, 0.21)	0.086	< 0.001	0.21 (0.13, 0.28)	0.101	< 0.001
Gender: male to female	−0.86 (−1.66, −0.07)	0.026	0.034	–	–	–
Age	–	–	–	-0.04 (−0.08, −0.01)	0.039	0.016
BMI	−0.03 (−0.12, 0.07)	0.005	0.586	–	–	–
TST	–	–	–	0.01 (0.00, 0.02)	0.031	0.006
Sleep efficiency	−0.04 (−0.07, −0.02)	0.030	0.001	−0.01 (−0.06, 0.04)	< 0.001	0.691
AHI	−0.01 (−0.03, 0.01)	0.004	0.220	–	–	–
Adjusted R^2	0.132			0.153		

Beta: an estimated regression coefficient for each variable with 95% confidence interval (95% CI).

The poor predictive abilities of both questionnaires for the PSG sleep disorders may be explained by the finding of multiple regression analyses that used scores from questionnaire as the dependent variable. Among the variables examined, including age, gender, BMI, the HADS score, TST, PLMI, AHI, and sleep efficiency, five (HADS score, gender, BMI, sleep efficiency, AHI) and four variables (HADS score, age, TST, and sleep efficiency) were selected at a threshold of $P < 0.10$ by univariate analyses for the PSQI and ESS, respectively. These findings are consistent in part with previous studies showing association between the PSQI and sleep efficiency, depression scale [6], or AHI [19] and association between the ESS and TST [19] or depression scale [20]. Among these screened variables, the variable with the most influence on PSQI and ESS scores was symptoms of depression and anxiety measured by the HADS. This finding is also consistent with previous studies, which have shown that PSQI and ESS scores are not or are only weakly associated with PSG measures, but are strongly associated with psychological symptoms, particularly, depression and anxiety [6,18–21]. The weak or absent associations of the PSQI and ESS scores with objectively defined sleep disorders might reflect the subjective nature of these measurements.

Strengths of the current study include a relatively large sample with a broad spectrum of patients typically seen in clinical practice, and the diagnosis of sleep disorders using a laboratory-based PSG. Nevertheless, our findings should be viewed with some degree of caution. First, the number of patients with RBD and narcolepsy were relatively small, which leads to relatively large CIs for the sensitivity, specificity, and AUC. Second, in the meta-analysis on the predictive ability of the ESS to identify OSA, widely used literature databases other than Medline, such as Embase, were not used, and a hand search of journals regarding sleep disorders was not conducted. This limitation means that all relevant articles may not have been included in the meta-analysis. However, we successfully extracted three articles that previous meta-analysis failed to find, which alleviates concern about comprehensiveness of the literature review. The last issue concerns the reliability of primary insomnia included in the other sleep disorders group. Because we did not examine psychiatric diagnoses by (semi-)structured interview, there is a possibility that we could not have distinguished primary insomnia from insomnia related to mental disorders. In fact, the HADS score was shown to be relatively high in the other sleep disorders group, mostly comprised of primary insomnia (Table 1) [26].

In summary, this study provides evidence for criterion validity of the PSQI and ESS using the diagnoses of OSA, PLMD, RBD, and narcolepsy as criterion standard. This study also includes and an up-to-date synthesis of the literature, including this study, regarding the accuracy of the ESS in screening for OSA. Given that neither the PSQI nor the ESS were designed to specifically screen for these sleep disorders, it was not surprising that the two questionnaires were of limited utility for the identification of OSA, PLMD, RBD, and narcolepsy. This finding might be explained by the fact that the PSQI and ESS scores are more influenced by psychological symptoms than PSG indices. Therefore, the PSQI and ESS should no longer be used as a screening or diagnostic instrument for PSG-defined sleep disorders. At maximum, both scales may provide information that contributes to the final diagnosis of sleep disorders.

Funding sources

This study was supported by the Grants-in-Aid for Scientific Research (Basic Research B: No. 24592578) from the Ministry of Education, Culture, Sports, Science and Technology in Japan and the Suzuken Memorial Foundation.

Conflict of interest

The ICMJE Uniform Disclosure Form for Potential Conflicts of Interest associated with this article can be viewed by clicking on the following link: http://dx.doi.org/10.1016/j.sleep.2013.12.015.

Acknowledgment

The authors would like to thank the participants in this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.sleep.2013.12.015.

References

[1] Moul DE, Hall M, Pilkonis PA, Buysse DJ. Self-report measures of insomnia in adults: rationales, choices, and needs. Sleep Med Rev 2004;8:177–98.
[2] Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28:193–213.
[3] Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991;14:540–5.
[4] Backhaus J, Junghanns K, Broocks A, Riemann D, Hohagen F. Test–retest reliability and validity of the Pittsburgh Sleep Quality Index in primary insomnia. J Psychosom Res 2002;53:737–40.
[5] Johns MW. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the Epworth Sleepiness Scale: failure of the MSLT as a gold standard. J Sleep Res 2000;9:5–11.
[6] Grandner MA, Kripke DF, Yoon I-Y, Youngstedt SD. Criterion validity of the Pittsburgh Sleep Quality Index: investigation in a non-clinical sample. Sleep Biol Rhythms 2006;4:129–39.
[7] Pouliot Z, Peters M, Neufeld H, Kryger MH. Using self-reported questionnaire data to prioritize OSA patients for polysomnography. Sleep 1997;20:232–6.
[8] Osman EZ, Osborne J, Hill PD, Lee BW. The Epworth Sleepiness Scale: can it be used for sleep apnoea screening among snorers? Clin Otolaryngol Allied Sci 1999;24:239–41.
[9] Sharkey KM, Machan JT, Tosi C, Roye GD, Harrington D, Millman RP. Predicting obstructive sleep apnea among women candidates for bariatric surgery. J Women's Health 2010;19:1833–41.
[10] Danzi-Soares NDJ, Genta PR, Nerbass FB, Pedrosa RP, Soares FSN, César LAM, et al. Obstructive sleep apnea is common among patients referred for coronary artery bypass grafting and can be diagnosed by portable monitoring. Coron Artery Dis 2012;23:31–8.
[11] Vana KD, Silva GE, Goldberg R. Predictive abilities of the STOP-Bang and Epworth Sleepiness Scale in identifying sleep clinic patients at high risk for obstructive sleep apnea. Res Nurs Health 2013;36:84–94.
[12] Albuquerque FN, Calvin AD, Sert Kuniyoshi FH, Konecny T, Lopez-Jimenez F, Pressman GS, et al. Sleep-disordered breathing and excessive daytime sleepiness in patients with atrial fibrillation. Chest 2012;141:967–73.
[13] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926–30.
[14] Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chron Dis 1985;38:27–36.
[15] Ross SD, Sheinhait I, Harrison KJ, Kvasz M, Connelly JE, Shea S, et al. Systematic review and meta-analysis of the literature regarding the diagnosis of sleep apnea. Sleep 2000;23:519–32.
[16] Ramachandran SK, Josephs L. A meta-analysis of clinical screening tests for obstructive sleep apnea. Anesthesiology 2009;110:928–39.
[17] Abrishami A, Khajehdehi A, Chung F. A systematic review of screening questionnaires for obstructive sleep apnea. Can J Anaesth 2010;57:423–38.
[18] Wells RD, Day RC, Carney RM, Freedland KE, Duntley SP. Depression predicts self-reported sleep quality in patients with obstructive sleep apnea. Psychosom Med 2004;66:692–7.
[19] Kezirian EJ, Harrison SL, Ancoli-Israel S, Redline S, Ensrud K, Claman DM, et al. Behavioral correlates of sleep-disordered breathing in older women. Sleep 2007;30:1181–8.
[20] Ishman SL, Cavey RM, Mettel TL, Gourin CG. Depression, sleepiness, and disease severity in patients with obstructive sleep apnea. Laryngoscope 2010;120:2331–5.
[21] Macey PM, Woo MA, Kumar R, Cross RL, Harper RM. Relationship between obstructive sleep apnea severity and sleep, depression and anxiety symptoms in newly-diagnosed patients. PloS One 2010;5:e10211.
[22] Doi Y, Minowa M, Uchiyama M, Okawa M. Development of the Japanese version of the Pittsburgh Sleep Quality Index. Jap J Psychiatry Treat 1998;13:755–63.

[23] Takegami M, Suzukamo Y, Wakita T, Noguchi H, Chin K, Kadotani H, et al. Development of a Japanese version of the Epworth Sleepiness Scale (JESS) based on item response theory. Sleep Med 2009;10:556–65.
[24] Kugaya A, Akechi T, Okuyama T, Okamura H, Uchitomi Y. Screening for psychological distress in Japanese cancer patients. Jap J Clin Oncol 1998;28:333–8.
[25] Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatry Scand 1983;67:361–70.
[26] Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res 2002;52:69–77.
[27] Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. (LA): Brain Information Sevice/Brain Research Institute; 1968.
[28] Iber C, Ancoli-israel S, Chesson A. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specification. 1st ed. Westchester (IL): American Academy of Sleep Medicine; 2007.
[29] American Association of Sleep Disorders. EEG arousals: scoring rules and examples: a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association. Sleep 1992;15:173–84.
[30] American Association of Sleep Disorders. Recording and scoring leg movements. The Atlas Task Force. Sleep 1993;16:748–59.
[31] American Academy of Sleep Medicine. International classification of sleep disorders, second edition (ICSD-2): diagnostic & coding manual. Westchester (IL): American Academy of Sleep Medicine; 2005.
[32] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
[33] Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667–76.
[34] Efron B, Tibshirani R. An introduction to the bootstrap. 1st ed. New York: Chapman & Hall/CRC; 1994.
[35] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
[36] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.
[37] Holling H, Bohning W, Bohning D. Meta-analysis of diagnostic studies based upon SROC-curves: a mixed model approach using the Lehmann family. Statist Model 2012;12:347–75.
[38] Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088–101.
[39] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. PROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform 2011;12:77.
[40] Canty A, Ripley B. Boot: bootstrap R (S-Plus) functions. 2012.
[41] Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Statist Softw 2010;36(3).
[42] Doebler P. Mada: meta-analysis of diagnostic accuracy; 2013.
[43] R Development CoreTeam. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
[44] Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev Veterin Med 2000;45:23–41.
[45] Sakuta K, Komada Y, Kagimura T, Okajima I, Nakamura M, Inoue Y. Factors associated with severity of daytime sleepiness and indications for initiating treatment in patients with periodic limb movements during sleep. Sleep Biol Rhythms 2012;10:187–94.
[46] Chung F, Yegneswaran B, Liao P, Chung S, Vairavanathan S, Islam S, et al. Validation of the Berlin questionnaire and American Society of Anesthesiologists checklist as screening tools for obstructive sleep apnea in surgical patients. Anesthesiology 2008;108:822–30.
[47] Boeve BF, Molano JR, Ferman TJ, Smith GE, Lin S-C, Bieniek K, et al. Validation of the Mayo Sleep Questionnaire to screen for REM sleep behavior disorder in an aging and dementia cohort. Sleep Med 2011;12:445–53.
[48] Hublin C, Kaprio J, Partinen M, Koskenvuo M, Heikkilä K. The Ullanlinna Narcolepsy Scale: validation of a measure of symptoms in the narcoleptic syndrome. J Sleep Res 1994;3:52–9.